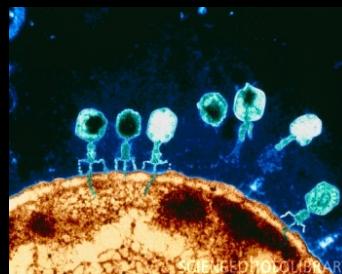
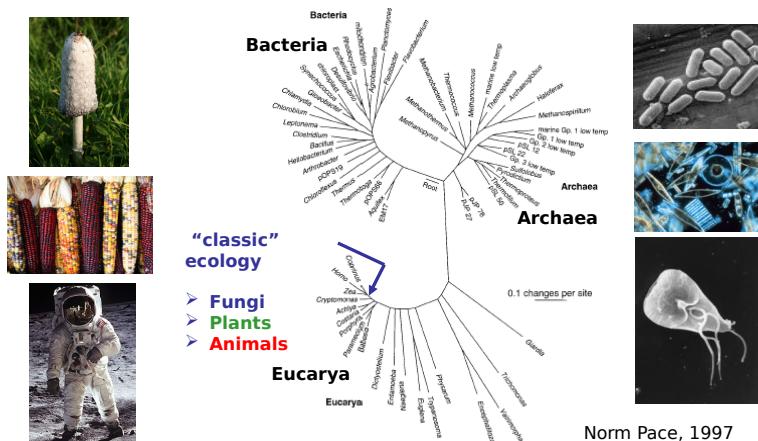


# Conceptos de filo-informática para investigación en genómica, ecología y evolución microbiana

Pablo Vinuesa, Centro de Ciencias Genómicas – UNAM  
vinuesa [at] ccg . unam . mx  
<http://www.ccg.unam.mx/~vinuesa/>

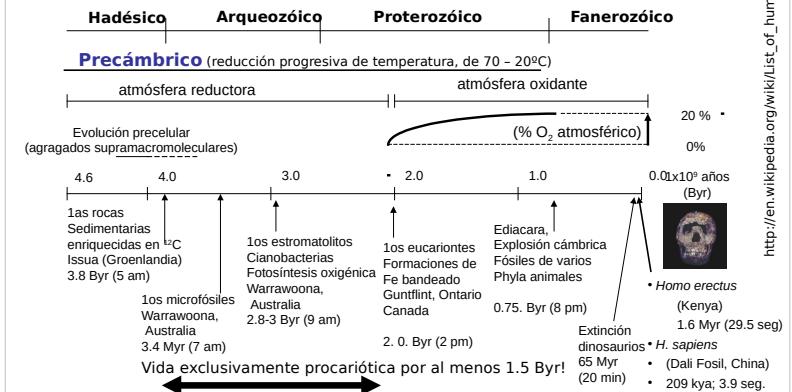


## Biodiversity = microbial diversity



## Evolución orgánica - La dimensión temporal

## **Historia de la tierra y de la vida**



## Parte I: Introducción a la Inferencia Filogenética

## Conceptos básicos:

- \* filogenia, evolución molecular y homología

## **Libros de referencia recomendados**

- Felsenstein, J., 2004. Inferring phylogenies. Sinauer Associates, INC., Sunderland, MA.

Futuyma, D.J. 2005. Evolution. Sinauer Associates, INC., Sunderland, MA.

Graur, D., Li, W.H., 2000. Fundamentals of Molecular Evolution. Sinauer Associates, Inc., Sunderland.

Nei, M., Kumar, S., 2000. Molecular Evolution and Phylogenetics. Oxford University Press, Inc., NY.

Page, R.D.M., Holmes, E.C., 1998. Molecular Evolution - A Phylogenetic Approach. Blackwell Science Ltd, Oxford.

Swofford, D.L., Olsen, G.J., Waddel, P.J., Hillis, D.M., 1996. Phylogenetic inference. In: Hillis, D.M., Moritz, C., Mable, B.K. (Eds.), Molecular Systematics. Sinauer Associates, Sunderland, MA, pp. 407-514. (Una revisión excelente del campo antes de aparecer los métodos Bayesianos)

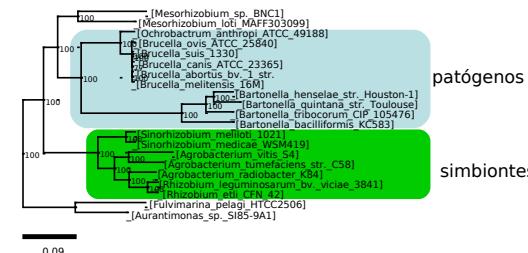
# Introducción a la Inferencia Filogenética – las especies y genomas microbianos

# Introducción a la filoinformática – pan-genómica y filogenómica. TIB2019, Junlio-Agosto 2019

## La relación entre filogenética y evolución molecular:

- La **filogenética** tiene por objetivo el **trazar la relación ancestro descendiente de los organismos** (árbol filogenético) a diferentes niveles taxonómicos, incluyendo el árbol universal, haciendo una reconstrucción de esta relación en base a diversos **caracteres homólogos**, tanto **morfológicos** como **moleculares**.

Las hipótesis filogenéticas resultantes son la base para hacer **predicciones** (inferencias) sobre propiedades biológicas de los grupos revelados por la filogenia mediante el mapeo de caracteres sobre la topología (hipótesis evolutiva). También proveen el contexto comparativo para poder inferir patrones de **evolución molecular**.

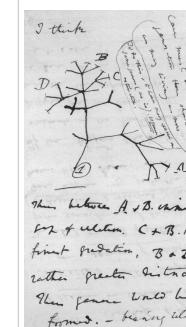


## Evolución de la filogenética como disciplina científica



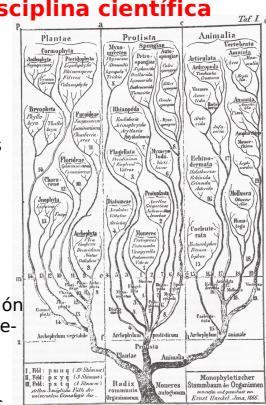
Los primeros intentos de reconstruir la historia filogenética estaban basados en pocos o ningún criterio objetivo.

Reflejaban las ideas o hipótesis plausibles generadas por expertos de grupos taxonómicos particulares.



I think  
Then follows A & B. Then  
C & D. Then  
fruit predation, B & D  
in other greater subtlety.  
Then follows L and he  
formed. - Henry Welles

La mayor parte de la 1a. mitad del SXIX los sistemáticos estaban más preocupados por el problema de definir a las especies biológicas, descubrir mecanismos de especiación y la variación geográfica de las especies, que en entender su filogenia.



No fue hasta los 40's y 50's que los esfuerzos de individuos como Walter Zimmermann y Willi Henning comenzaron a definir métodos objetivos para reconstruir filogenias en base a caracteres compartidos entre organismos fósiles y contemporáneos.

## ¿Porqué estudiar filogenética y evolución molecular?

Corolario I:

"Nothing in biology makes sense except in the light of evolution"

- Theodosius Dobzhanski, 1973  
(The American Biology Teacher 35:125)

Corolario II:

"Nothing in evolutionary biology makes sense except in the light of a phylogeny"

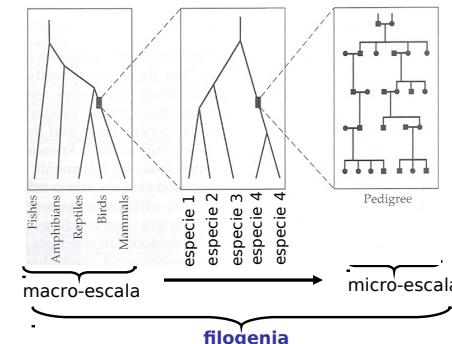
- Jeff Palmer, Douglas Soltis, Mark Chase, 2004  
(American J. Botany 91: 1437-1445)



## El concepto de filogenia y homología: definiciones básicas

"**The stream of heredity makes phylogeny**; in a sense, it is phylogeny. Complete genetic analysis would provide the most priceless data for the mapping of this stream".

G.G. Simpson (1945)



**Filogenia:** historia evolutiva del flujo hereditario a distintos niveles evolutivos/temporales, desde la genealogía de genes en poblaciones (micro-escala; dominio de la genética de poblaciones) hasta el árbol universal (macro-escala)

## El concepto de filogenia y homología: definiciones básicas

**Homología:** es la relación entre dos caracteres que han descendido generalmente con modificación, de un ancestro común. Estrictamente se refiere a ancestría común inferida.

**Analogía:** es la relación existente entre dos caracteres cuando éstos, aún siendo similares, han heredados convergentemente a partir de caracteres ancestrales no relacionados en términos genealógicos.

**Cenancesto:** del inglés (cenancestor), es el ancestro común más reciente de los taxa bajo consideración.

## El concepto de homología: definiciones básicas

Dado que filogenia es “el flujo de la herencia”, **sólo los caracteres genéticos o heredables son informativos desde una perspectiva genealógica**

**Caracteres y estados de carácter.** Los evolucionistas distinguen entre caracteres, como por ejemplo los amino ácidos, y sus estados de carácter, como pueden ser gly o trp. La homología reside en los caracteres, no en sus estados!!!

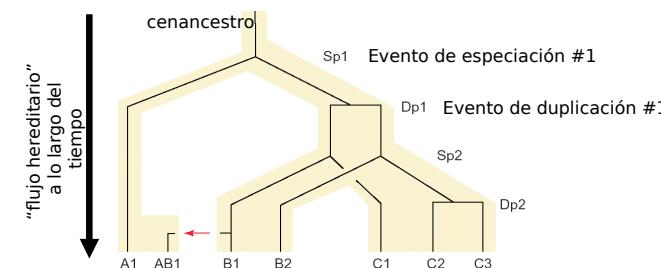
El reconocimiento de la condición de homología entre caracteres. **La homología no es una cualidad cuantitativa.** Sólo hay dos condiciones posibles: ser o no homólogo. No se es más o menos homólogo. Es como el embarazo. Se está o no se está en dicho estado y se es o no homólogo.

Por tanto, para cuantificar el parecido entre un par de secuencias homólogas se dice que presentan globalmente un 70% y 95% de **identidad y similitud**, respectivamente. (no existe algo como 95% de homología).

**El concepto de homología es simplemente una abstracción sobre la relación entre caracteres, sobre su ascendencia común,** relación que es indispensable determinar para poder hacer reconstrucciones filogenéticas que reflejen la historia del “flujo de la herencia”.

## El concepto de homología: definiciones básicas

### Subtipos de homología: ortología, paralogía y xenología



**ortología:** relación entre secuencias en la que la divergencia acontece tras un evento de especiación. El ancestro común es el cenancestro. La filogenia recuperada de estas secuencias refleja la filogenia de las especies.

**paralogía:** condición evolutiva en la que la divergencia observada acontece tras un evento de duplicación génica. La mezcla de ortólogos y parálogos en un mismo análisis filogenético recupera la filogenia correcta de los genes pero no necesariamente la de los organismos o taxa.

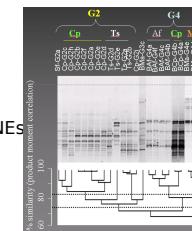
**xenología:** relación entre secuencias dada por un evento de transferencia horizontal entre linajes. Distorsiona fuertemente la filogenia de las especies.

## Marcadores moleculares usados en filogenética y evolución molecular

### Polimorfismos de DNA y proteínas

#### I) Marcadores dominantes ( $\neq$ secuencias)

- RFLPs
- Fingerprints genómicos (AFLPs, RAPDs, Rep-PCR, SINEs, SSRPs, NSNPs ...)
- Análisis multilocus de isoenzimas
- etc ...



Los datos moleculares revelan información genética. Sólo datos con una base genética son de interés en filogenética y evolución. De ahí que los marcadores moleculares son generalmente los favorecidos para hacer inferencias filogenéticas y evolutivas a distintos niveles taxonómicos.

Los caracteres fenotípicos muchas veces tienen una base genética menos clara y están gobernados por las interacciones de muchos genes con el ambiente. Muchos fenotipos presentan gran plasticidad, es decir, que un mismo genotipo puede presentar una gradación de fenotipos. Esta variación fenotípica puede confundir las verdaderas relaciones filogenéticas y determinación de parentescos.

El uso de protocolos de PCR permite acceder a todo el mundo biológico para scrutinios genéticos

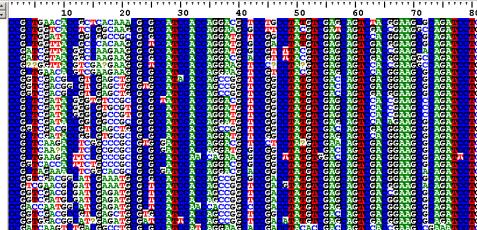
Los métodos moleculares permiten una fácil y robusta distinción entre homología y analogía y permiten hacer comparaciones de divergencia evolutiva usando métricos universales

# Introducción a la Inferencia Filogenética – las especies y genomas microbianos

# Introducción a la filoinformática – pan-genómica y filogenómica. TIB2019, Junlio-Agosto 2019

Marcadores moleculares usados en filogenética y evolución molecular

## II) Secuencias moleculares DNA/proteína



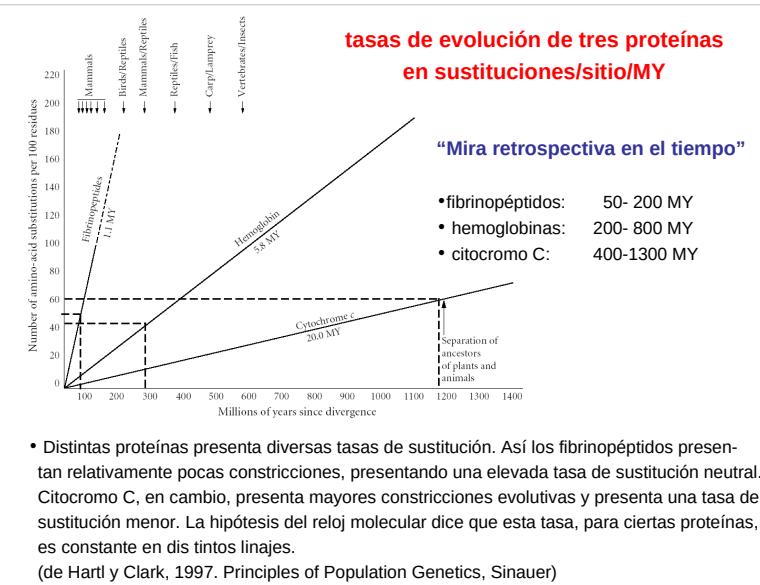
- La premisa fundamental en evol. molec. es que en dichas secuencias se encuentra escrita una buena parte de su historia evolutiva.
  - Secuencias de DNA representan el “nivel anatómico” más fino de un organismo
  - Buena parte de la biología moderna tiene por objetivo revelar la información contenida en secuencias moleculares
  - Para inferir la historia de relaciones de ancestría entre un conjunto de secuencias homólogas hemos de **determinar las correspondencias de homología entre los caracteres** haciendo un **alineamiento múltiple de las secuencias**

## Selección de marcadores adecuados para hacer inferencias evolutivas a distintos niveles de profundidad filogenética

## Restricciones funcionales vs. tasas de sustitución:

- Existe gran variabilidad en la tasa de sustitución entre genes y dominios génicos:
    - intrones vs. exones
    - regiones codificadoras vs. regiones intergénicas o pseudogenes
    - residuos catalíticos vs. no catalíticos, dominios estructurales vs. no estructurales
    - 3as. posiciones vs. 1as y 2as en codones de secuencias codificadoras,
    - asas vs. orquillas en rRNAs y tRNAs ...
  - Existen genes de evolución muy rápida o muy lenta:
    - fibrinopéptidos evolucionan una tasa x900 > a la de ubiquitina y x20 > citocromo C
    - genes de HIV evolucionan a  $\times 10^6$  veces la tasa de un gen humano promedio!
  - **Tasas de evolución y la teoría neutral de evolución molecular:**

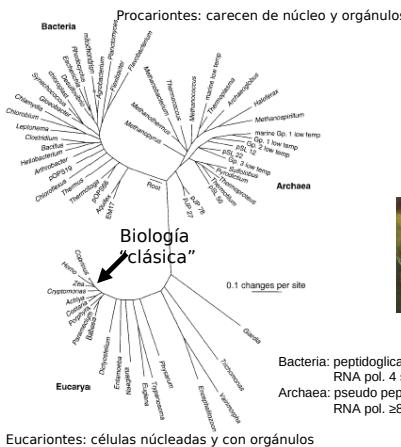
el reloj molecular, calibración y datación de eventos de especiación/extinción de linajes y pandemias ...



Aplicaciones y predicciones filogenéticas (I)

- Elucidación del árbol universal, sistemática bacteriana y la identificación/clasificación de microorganismos ambientales (cultivables y NO CULTIVABLES > 90-99%)

## *rrs: un marcador lento*



Bacteria: peptidoglicano; lípidos de membrana son ésteres de glicerol;  
RNA pol. 4 subunidades; formilmetionina como aa de inicio ...

Archaea: pseudo peptidoglicano; lípidos de membrana son ésteres de glicerol;  
RNA pol. >8 subunidades; metionina como aa de inicio ...

**Aplicaciones y predicciones filogenéticas (II):**  
Evidencia molecular de transmisión de HIV-1 en un caso criminal usando genes de evol. rápida

Un gastroenterólogo fue acusado del intento de asesinato en 2º grado de su novia mediante inyección de sangre contaminada con HIV-1.

Este estudio representa el primer caso en el que reconstrucciones filogenéticas de secuencias (paciente P, víctima V y controles LA de portadores en la población) fueron admitidas en una corte criminal en EUA.

Las filogenias de RT y de env mostraron que las secuencias de la V comparten ancestría directa en forma de paralogía con las de una P del gastroenterólogo.

Análisis de posiciones de codones de la RT de la V revelaron genotipos consistentes con mutaciones que confieren AZTR, similares a las presentadas en la P.

El establecimiento a priori de la P y V como posible par de transmisión del HIV-1 representó una clara hipótesis para ser evaluada en marcos de estadística filogenética.

Ref: Metzker et al. 2002. PNAS 99:14292-142976

Filogenias del gen RT basadas en secuencias de la V, la P y LA, obtenidas por dos labs. independientes.  
a) Baylor College of Medicine, Houston, TX (BMC)  
b) Dpt. Ecology and Evol. Biol., Univ. Michigan (MIC)

## Part 2 – Introduction to microbial molecular systematics and taxonomy

How are bacterial species defined

Gene trees vs species tree

MLSA, MLST and genomic taxonomy

Selection of molecular markers for microbial systematics

Prokaryotic sex and its consequences for microbial taxonomy and systematics

Two case studies: from MLSA to phylogenomics

**¿What are bacterial species?**

- an official taxonomic definition:  
the phylogenetic species concept

Stackebrandt E, Frederiksen W, Garrity GM, et al. (2002)  
**Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology.** *Int. J. Syst. Evol. Microbiol.* **52**, 1043-1047

"A species is a **category** that circumscribes a (preferably) **genomically coherent group** of individual isolates/strains **sharing a high degree of similarity** in (many) independent features, comparatively tested under highly standardized conditions"

- DNA:DNA homology > 70% and > 97% similarity in **16S rRNA sequences**
- A **polyphasic approach** to determine the degree of genotypic and phenotypic similarity and to "search for a diagnosable and discriminative phenotypic property"
- **ad hoc** concept for prokaryotes
- **tipological conception**; species as static entities with a "**typus strain**" being representative of the species (> 90% of the spp. described in the IJSEM in the past 3 years are based on a single isolate!!!)
- Does not consider actual knowledge about bacterial **genomics** and evolutionary biology

## The evolutionist's way to look at species

1.- sample large collections of strains from related species recovered from different demes

2.- Molecular evolutionary analyses based on sequences for multiple loci (core and auxiliary genes):

a) inference of **multiple gene phylogenies** to identify congruent and significantly supported evolutionary lineages (distinct species?)

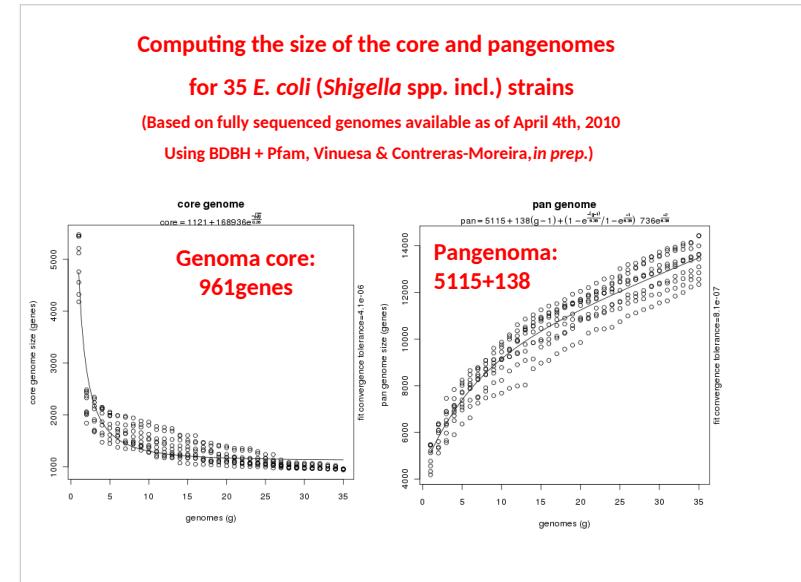
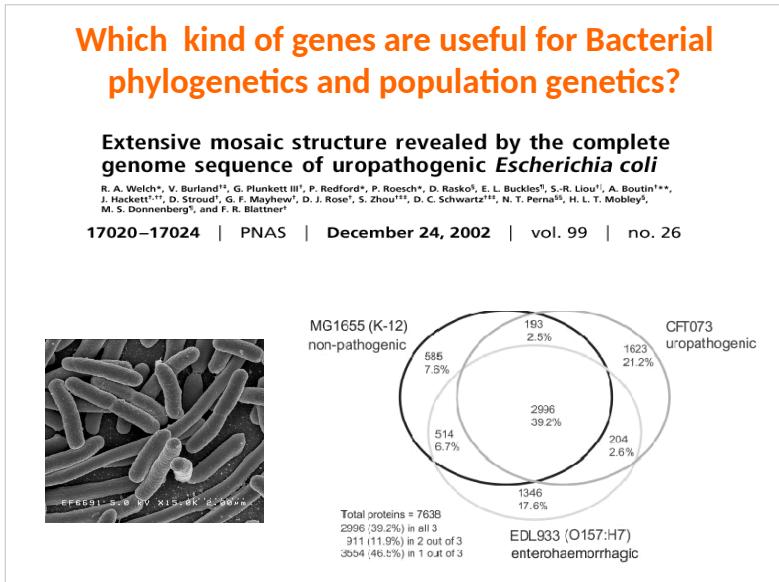
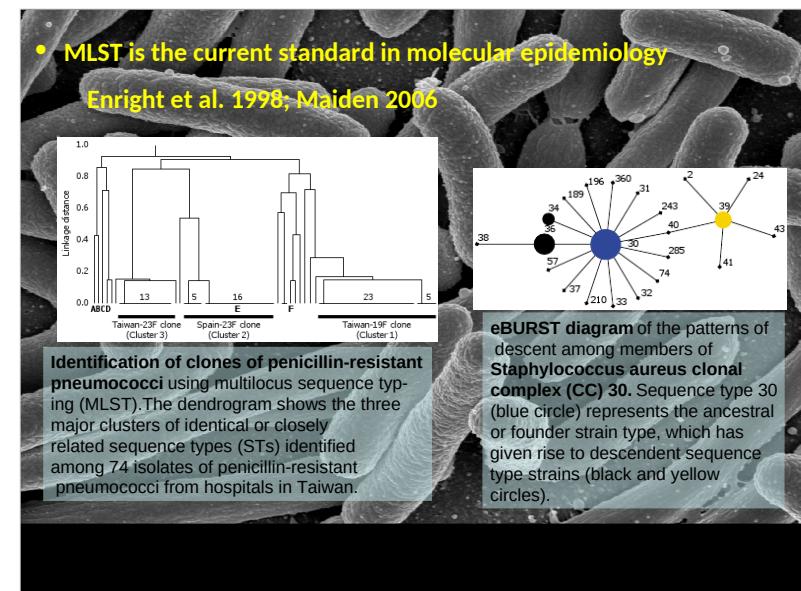
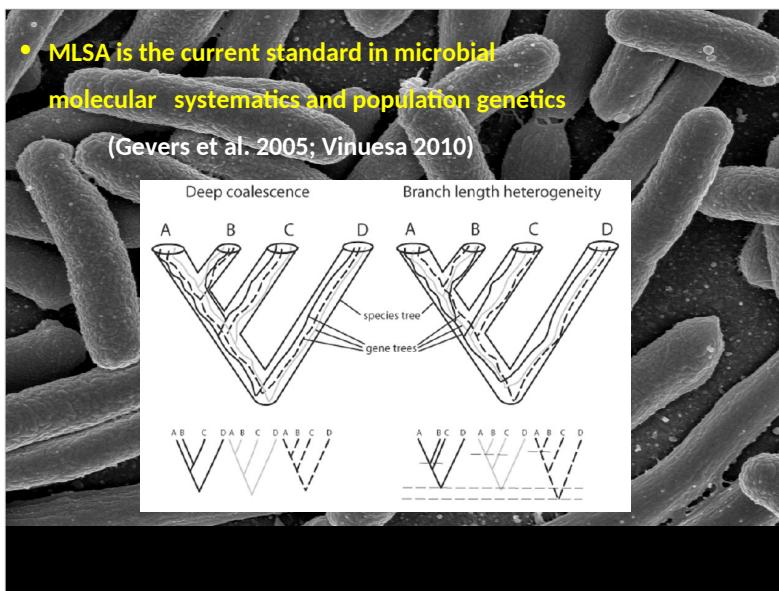
b) **DNA polymorphism analyses** of the strains grouped in different phylogenetic clades

- does significant **genetic differentiation** exist between members of different clades?
- which are the **evolutionary forces** shaping the **genetic structure of populations** and providing them with **internal cohesion**?

3.- The emerging field of genomic taxonomy:  
study of core- and pan-genome phylogenies,  
computing of overall genome similarities (cgANI)

# Introducción a la Inferencia Filogenética – las especies y genomas microbianos

# Introducción a la filoinformática – pan-genómica y filogenómica. TIB2019, Junlio-Agosto 2019





### Natural history of rhizobia - N<sub>2</sub>-fixing legume nodule microsymbionts

**Isolation of symbiotic rhizobia from endemic Canarian genistoid legumes**

Nodules      CLSM of GFP-expressing rhizobia within a nodule  
500 µm

- Field-collected nodules
- Nodules induced in the laboratory on particular trap plants using native soils as the inoculum

*Chamaecytisus proliferus* (Tagasaste)

### Natural history of Rhizobia from the Canary Islands

locations of sampled populations

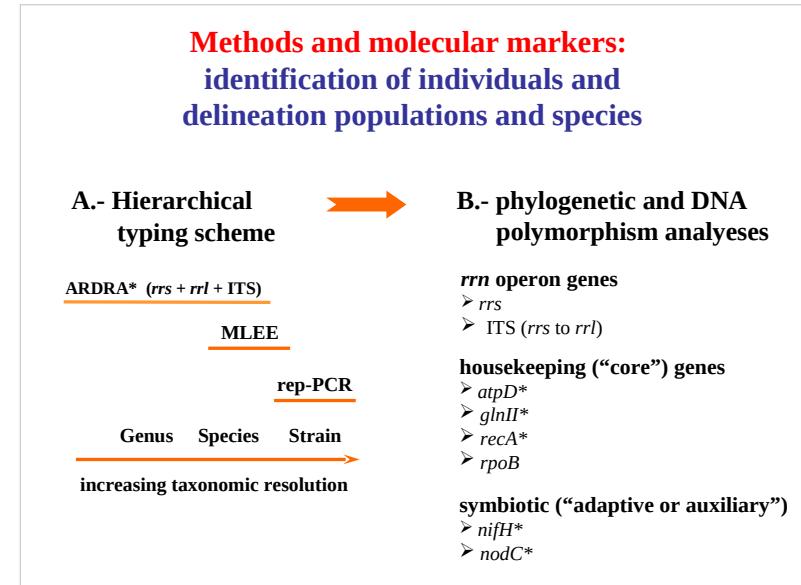
Adenocarpus foliolosus ("Codeso")

- Can bacterial species boundaries be recognized?
- How many species nodulate ECGLs?
- Which forces shape their population genetic structures?
- geographically structured populations?

**Bradyrhizobium strains (n=315)**  
(23 host genera in 11 tribes, from 24 countries on 4 continents)

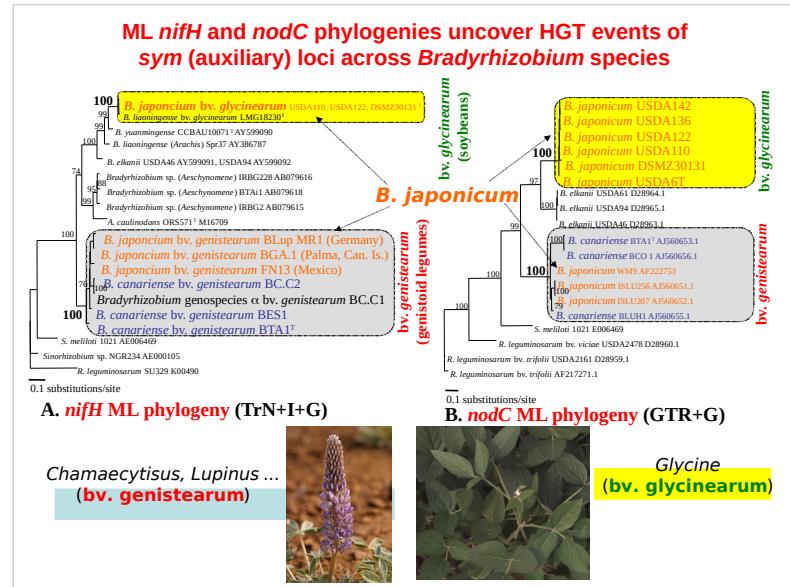
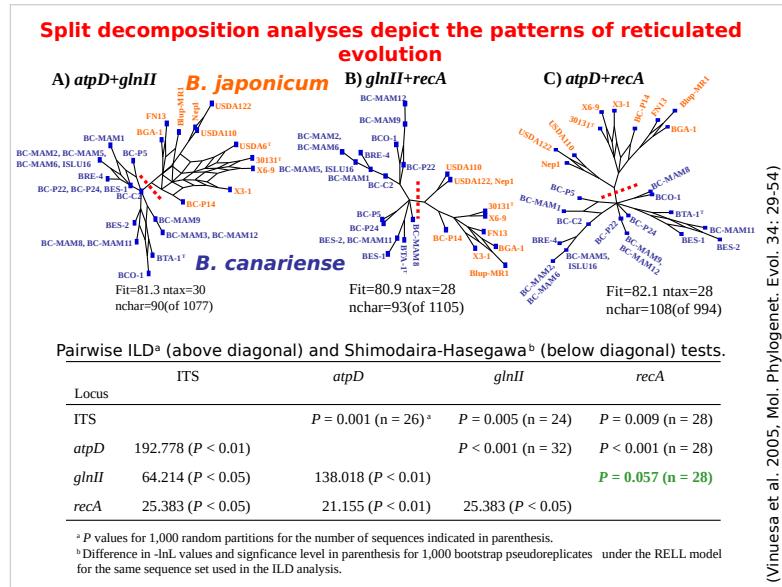
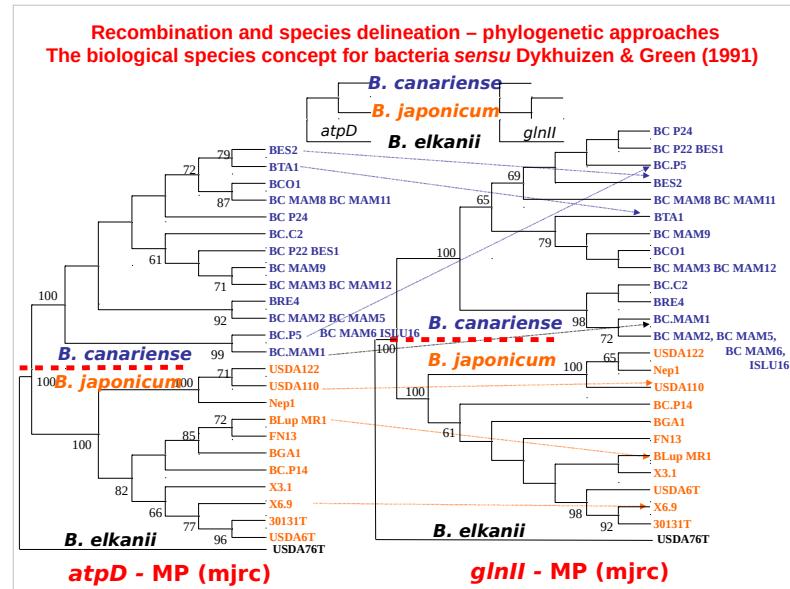
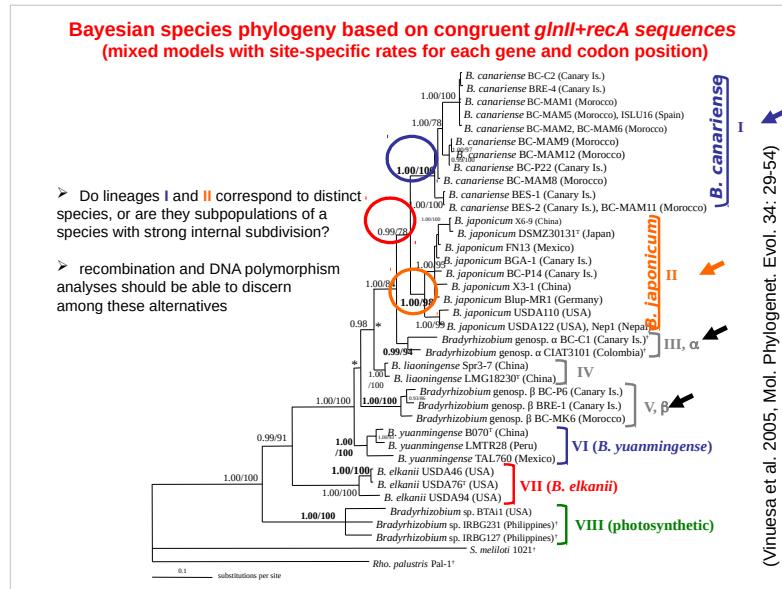
Ref. isolates for phylogenetic analyses ECGL isolates

Collection	nº of isolates	hosts	geographic origin	ref/source
Canarian isolates	126	<i>Chamaecytisus</i> , <i>Spartocytisus</i> , <i>Teline</i>	Gran Canaria, La Palma, Gomera Tenerife	Vinuesa et al., 1998. AEM 64:2096–2104 Vinuesa, Mol. Phylogenet. Evol. 2005 Vinuesa, Int. J. Syst. Evol. Microbiol. 2005
Maroccan isolates	24	<i>Chamaecytisus</i>	M. Kenitra, Mammora	
CIAT	20	<i>Alysicarpus</i> , <i>Calopogonium</i> , <i>Centrosema</i> , <i>Desmodium</i> , <i>Macrotyloma</i> , <i>Pueraria</i> , <i>Stylosanthes</i>	Australia, Brazil, Colombia, Hawaii, Kenya, Mexico, Thailand, Malaysia, Peru, Zimbabwe	Centro Internacional de Agricultura Tropical, Cali, Colombia
Chinese isolates from <i>G. max</i>	3	<i>Glycine</i>	China	Holger Blasum
CIFN	2	<i>Lupinus</i>	Mexico	Barrera et al., 1997. Int. J. Syst. Bacteriol. 47:1086–1091. So et al., 1994. Int. J. Syst. Bacteriol. 44:392–403
IRRI	10	<i>Aeschynomene</i> , <i>Acacia</i> , <i>Crotalaria</i> , <i>Glycine</i> , <i>Indigofera</i> , <i>Macrotyloma</i> , <i>Vigna</i>	Hawaii, India, Japan, Kenya, Mexico, Philippines, Zimbabwe	
HAMBI	6	<i>Arachis</i>	Hawaii, China	Zhang et al., 1991. Int. J. Syst. Bacteriol. 41:104–113 K. Lüttge, Univ. Hohenheim M. Sicardi and M. L. Izquierdo-Mayoral P. Vinuesa, Philipps-Universität Marburg
IVIC	8	<i>Glycine</i> , <i>Vigna</i>	Venezuela	
BOTAMAR	6	<i>Glycine</i> , <i>Lupinus</i>	USA, Germany	
USDA/ARS	4	<i>Glycine</i>	USA	P. van Berkum, USDA/ARS
LMG	1	<i>Glycine</i>	China	A. Willems, U. Gent
DSMZ	1	<i>Glycine</i>	China	
CIFN	5	<i>Phaseolus lunatus</i>	Peru	E. Ormeño, CIFN-UNAM
IPN	1	<i>Lespidea</i>	China	E.T. Wang, CB-IPN
LMG	1	<i>Glycine</i>	China	A. Willems, U. Gent
BOTAMAR	80	<i>Glycine</i>	India, Myanmar, Nepal, Vietnam	Vinuesa, Rojas-Jiménez, Mahna, Moe, Prasad, Babu, Thierfelder & Werner



# Introducción a la Inferencia Filogenética – las especies y genomas microbianos

# Introducción a la filoinformática – pan-genómica y filogenómica. TIB2019, Junlio-Agosto 2019



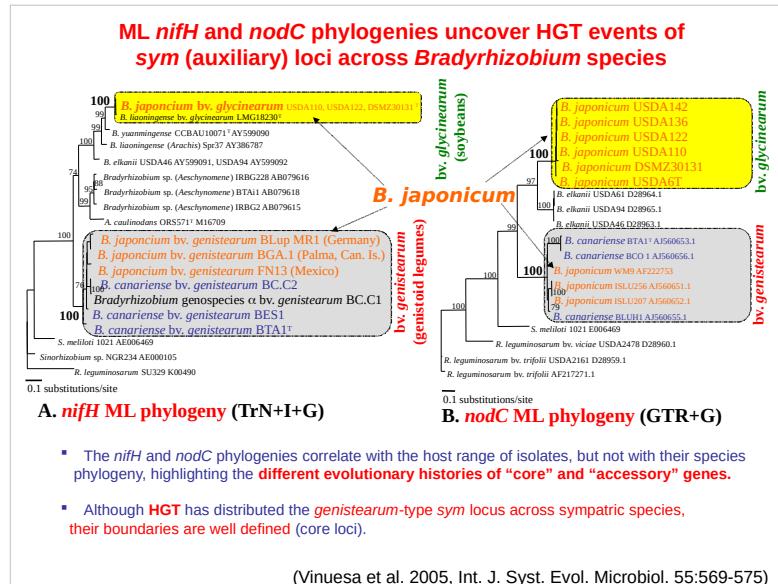
<sup>a</sup> P values for 1,000 random partitions for the number of sequences indicated in parentheses.

<sup>b</sup>Difference in -lnL values and significance level in parenthesis for 1,000 bootstrap pseudoreplicates under the RELL model for the same sequence set used in the ILD analysis.

© Pablo Vinuesa 2019,  
vinuesa{at}ccg{dot}unam{dot}mx  
<http://www.ccg.unam.mx/~vinuesa/>

# Introducción a la Inferencia Filogenética – las especies y genomas microbianos

# Introducción a la filoinformática – pan-genómica y filogenómica. TIB2019, Junlio-Agosto 2019



## Analysis of DNA polymorphisms: genetic differentiation and gene flow estimates for populations of *B. canariense* and its sister species *B. japonicum*

Table 5. Genetic differentiation and gene flow estimates

Gene / populations	No. Fix		Genetic differentiation		Gene flow		
	Dif. <sup>a</sup>	$\chi^2$ (df) <sup>b</sup>	P <sup>c</sup>	$K_{ST}^{*d}$	P <sup>e</sup>	$F_{ST}^{f}$	Nm <sup>g</sup>
atpD							
<i>B. canariense</i> insular vs. continental	0	13.333 (14)	0.5005 (ns)	0.00135	0.4151 (ns)	0.01458	33.78
<i>B. japonicum</i> vs. <i>B. canariense</i>	11	28.000 (17)	0.0449	0.20326	0.0000	0.57751	0.37
glbII							
<i>B. canariense</i> insular vs. continental	0	11.200 (8)	0.1906 (ns)	0.05014	0.1000 (ns)	0.09391	4.82
<i>B. japonicum</i> vs. <i>B. canariense</i>	9	28.000 (17)	0.0449	0.21687	0.0000	0.61930	0.31
recA							
<i>B. canariense</i> insular vs. continental	0	9.333 (10)	0.5008 (ns)	0.00107	0.3901 (ns)	0.04188	11.44
<i>B. japonicum</i> vs. <i>B. canariense</i>	10	28.000 (19)	0.0834 (ns)	0.16635	0.0000	0.58114	0.36

<sup>a</sup> Number of fixed differences between populations.

<sup>b</sup> Haplotype based statistic (Hudson et al. 1992a), degrees of freedom are indicated in parenthesis.

<sup>c</sup> Probability of rejecting the null hypothesis that the two populations are not genetically differentiated, based on the critical values from the  $\chi^2$  distribution.

<sup>d</sup> Sequence based statistic described in Hudson et al. (1992a).

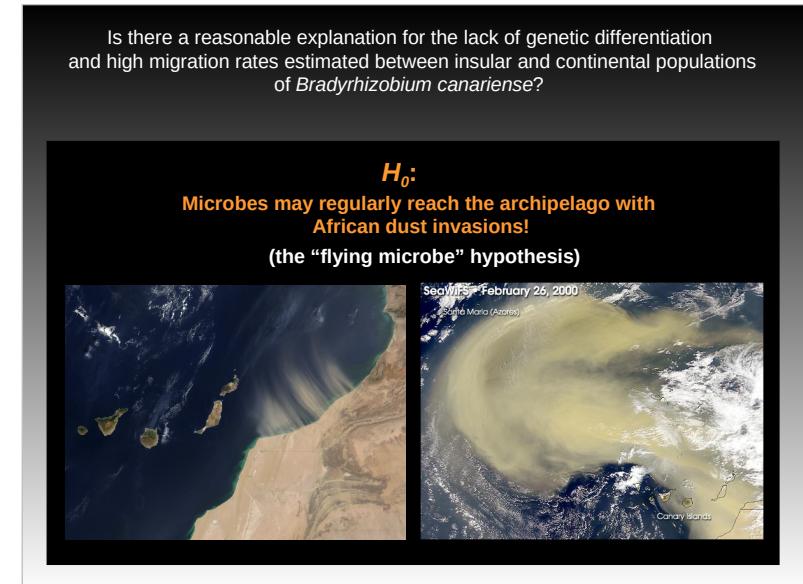
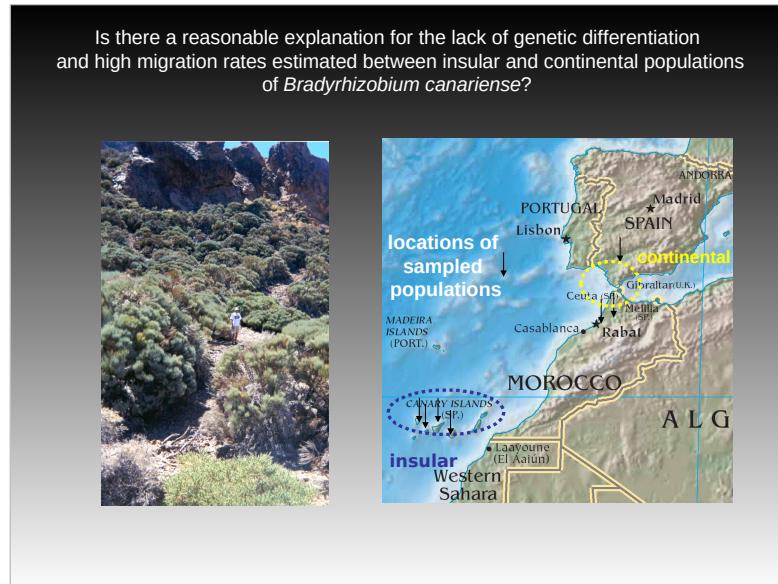
<sup>e</sup> Probability obtained by the permutation test (Hudson et al. 1992a) with 1000 replicates.

<sup>f</sup> Sequence based estimate described in Hudson et al. (1992a).

<sup>g</sup> Effective number of migrants.

ns, not significant.

(Vinuesa et al. 2005, Mol. Phylogenet. Evol. 34: 29-54)



# Introducción a la Inferencia Filogenética – las especies y genomas microbianos

# Introducción a la filoinformática – pan-genómica y filogenómica. TIB2019, Junlio-Agosto 2019

